第 45 卷第 2 期
2013 年 4 月

南 京 航 空 航 天 大 学 学 报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 45 No. 2
Apr. 2013

基于奇异值分解的致病基因挖掘算法

张焕萍[1,2]　尹佟明[2]　郑建冬[1]

(1. 南京林业大学机电学院，南京，210037；
2. 南京林业大学林木遗传与生物技术省部共建教育部重点实验室，南京，210037)

摘要：为从复杂的基因芯片表达数据中有效地挖掘致病基因，提出了基于奇异值分解(Singular value decomposition，SVD)的致病基因挖掘算法——Logistis 回归奇异值分解(Logistic regression SVD，LRSVD)，针对奇异值方差评估特征模式的不足，LRSVD 算法用 Logistic 回归系数代替方差评估每一个特征模式对分类的作用大小，进一步提出利用基因内积评估每一条基因的分类能力，建立了特征模式与原始基因表达数据之间的线性映射关系，并按基因内积大小排序，选择出对样本分类能力高的基因子集。将 LRSVD 算法应用于实际基因表达数据，实验结果表明，LRSVD 算法能有效挖掘出与疾病相关的基因子集。

关键词：奇异值分解；致病基因；Logistic 回归

中图分类号：TP181；Q811.4　　**文献标志码**：A　　**文章编号**：1005-2615(2013)02-0277-06

Disease Gene Identification Based on Singular Value Decomposition

Zhang Huanping[1,2], *Yin Tongming*[2], *Zheng Jiandong*[1]

(1. College of Mechanical and Electrical Engineering, Nanjing Forestry University, Nanjing, 210037, China;
2. Key Laboratory of Forest Genetics and Biotechnology Ministry of Education,
Nanjing Forestry University, Nanjing, 210037, China)

Abstract: To efficiently identify disease genes from gene expression data, an improved method Logistic regression singular value decomposition (LRSVD), based on singular value decomposition is proposed to find the genes associated with disease. LRSVD evaluates the contribution of each eigengene to the classifying accuracy by regression coefficients of Logistic regression instead of the variance. The inner-product (IP) is proposed to evaluate the discriminative power of each gene. The linear mapping relationship is established between eigengenes and the original gene expression data. Then the genes are ranked by the corresponding IP value and the optimal gene subset with high discriminative power of sample classification is identified. The obtained results on real gene expression data indicate that LRSVD has the ability of effective disease gene identification with high classifying accuracy.

Key words: singular value decomposition; disease genes; Logistic regression

DNA 微阵列技术可在一次实验中同时测定成千上万条基因的表达水平，已成为致病基因研究的重要途径。致病基因挖掘就是从海量基因中筛选出与疾病相关的致病基因或特征基因子集，使其对不同疾病样本的可分性最大，从而可以作为诊断基因，用于疾病的临床诊断。合理的致病基因子集不仅能提高分类效率，找到疾病亚型，使基因诊断治疗、药物研制等更有针对性。

基因表达数据具有基因数量多、样本数量少等特点，且基因之间的相互关系非常复杂，从基因表

基金项目：林业公益性行业科研专项重大(201304102)资助项目；国家自然科学基金(31270711)资助项目；长江学者和创新团队发展计划资助项目。

收稿日期：2012-08-24；**修订日期**：2012-10-14

通信作者：张焕萍，女，博士，讲师，1975 年生，E-mail：nuaazhp@126.com。

达数据中挖掘致病基因有一定难度。究竟选择多少个特征基因以及采用什么原则来选择这些基因，并没有统一的方法。奇异值分解(Singular value decomposition, SVD)对大型基因表达数据矩阵有很高的计算效率，并能检测出基因表达数据中微弱的表达模式，得到的特征模式具有一定的生物学意义[1-3]。奇异值分解在基因表达数据处理分析方面已有很多应用，如时间序列的基因表达数据分析[4]、用于缺失值弥补的 SVDimpute 算法[5]、基因聚类[6-8]、生物数据降维等[9]，但是对于基因样本分类及特征基因选择，并没有太多相关研究。本文提出了改进的基于 SVD 的特征基因选择算法——Logistic 回归 SVD(Logistic regression SVD, LRSVD)，算法首先对基因表达数据矩阵进行 SVD，得到矩阵中的特征模式，针对奇异值方差评估特征模式的不足，用 Logistic 回归系数来评估每个特征模式对分类的贡献大小，用基因内积表示每条基因在不同特征模式上的映射，并按内积大小对基因排序，多次迭代后得到最高分类准确率的最佳特征基因子集。将本文提出的算法应用于实际生物数据，实验结果表明，LRSVD 算法能有效挖掘出与疾病相关的基因子集。

1 LRSVD 算法原理

奇异值分解计算效率高，是线性代数中重要的矩阵分解，在信号处理、统计学、生物数据处理等领域有重要应用。基因表达数据矩阵具有基因数量多、样本数量少、高噪声等特点，其奇异值分解又不同于一般矩阵。

1.1 基因表达数据奇异值分解

用 $\boldsymbol{X}_{m\times n}=(x_{ij})_{m\times n}$ 表示基因表达数据矩阵，x_{ij} 表示第 i 条基因在第 j 个样本下的表达值，通常 $m\gg n$ 且矩阵的秩 $r=\min(m,n)$，矩阵的行表示基因在不同样本中的表达值，称为基因表达谱；矩阵的列表示样本条件下不同基因的表达值，称为样本表达谱。$\boldsymbol{X}_{m\times n}$ 奇异值分解为

$$\boldsymbol{X}_{m\times n}=\boldsymbol{U}_{m\times n}\boldsymbol{S}_{n\times n}\boldsymbol{V}_{n\times n}^{\mathrm{T}} \tag{1}$$

矩阵 $\boldsymbol{U}$ 的列 $\boldsymbol{u}_k$ 为左奇异向量，形成正交基，又称为基因系数向量或特征谱，是样本表达谱的线性组合；矩阵 $\boldsymbol{V}^{\mathrm{T}}$ 的行 $\boldsymbol{v}_k$ 是正交的右奇异向量，又称为特征模式，是基因表达谱的线性组合；$\boldsymbol{S}$ 是对角矩阵，对角线上非零元素 s_k 称为奇异值，又称为模式幅值，奇异值按降序排列，每一个奇异值对应一个左奇异向量 $\boldsymbol{u}_k$ 和右奇异向量 $\boldsymbol{v}_k$。矩阵 $\boldsymbol{X}_{m\times n}$ 可分解成：$\boldsymbol{X}_{m\times n}=\sum_{k=1}^{r}\boldsymbol{u}_k\,s_k\boldsymbol{v}_k^{\mathrm{T}}$，即奇异向量外积后的加权和，权重即是非零的奇异值。$\boldsymbol{X}^{(l)}\approx\sum_{k=1}^{l}\boldsymbol{u}_k\,s_k\boldsymbol{v}_k^{\mathrm{T}}$ 是矩阵 $\boldsymbol{X}$ 的近似矩阵，且秩为 $l(l<r)$。矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{X}^{(l)}$ 的差等于 $\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}-x_{ij}^{(l)}|^2$，当 $\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}-x_{ij}^{(l)}|^2$ 取最小值时，$\boldsymbol{X}^{(l)}$ 是降秩矩阵中 $\boldsymbol{X}$ 的最佳近似矩阵。选取不同 l 值，可得到不同的近似矩阵。奇异值的平方 s_k^2 与所对应的奇异向量的方差成正比，通常用奇异值方差贡献率 δ_k 来评判第 k 个奇异向量包含原数据的信息比例

$$\delta_k=s_k^2/\sum_{k=1}^{n}s_k^2 \tag{2}$$

1.2 特征模式数量选取

对基因表达数据矩阵 $\boldsymbol{X}_{m\times n}$ 进行奇异值分解，得到近似矩阵 $\boldsymbol{X}^{(l)}\approx\sum_{k=1}^{l}\boldsymbol{u}_k\,s_k\boldsymbol{v}_k^{\mathrm{T}}$，$\boldsymbol{X}^{(l)}$ 包含了原数据矩阵中的主要信息，用 $\boldsymbol{X}^{(l)}$ 代替原数据矩阵进行分析，可消除原数据中的噪声，降低数据维数。图 1 是近似矩阵的图形表示[2]。

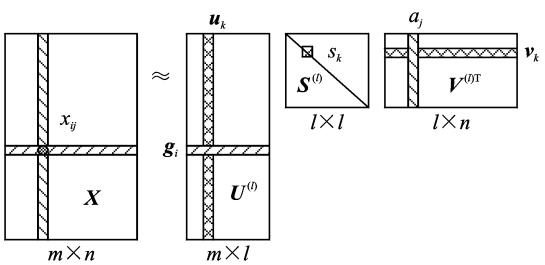

图 1 基因表达数据矩阵 SVD 的图形表示

特征模式 $\boldsymbol{v}_k$ 按对应奇异值大小降序排列，每一个特征模式代表一个投影方向，即包含在原基因表达数据中的模式，每个特征模式包含的信息量可用方差贡献率 δ_k 表示。基因系数向量 $\boldsymbol{u}_k=(u_{1k},u_{2k},\cdots,u_{mk})^{\mathrm{T}}$ 表示 m 条基因在第 k 个投影方向的投影大小，如 u_{1k} 表示第一条基因在第 k 个投影方向的投影，u_{1k} 数值越大，其投影越大；$\boldsymbol{g}_i=(g_{i1},g_{i2},\cdots,g_{il})$ 表示第 i 条基因分别在 l 个特征模式上的投影大小；$\boldsymbol{a}_j$ 表示压缩后的样本表达谱。

合理选取奇异值数量，用近似矩阵 $\boldsymbol{X}^{(l)}$ 代替原数据矩阵 $\boldsymbol{X}_{m\times n}$，可消除噪声及无关信息。若奇异值数量选取太少，会丢失包含在原数据中的信息，奇异值数量选取过多，则起不到降维和消噪的作用。奇异值数量选取方法与原始数据类型、数据来源有很大关系，对于多重共线性且噪声较低的数据，取前一至二个奇异值就足够了；基因表达数据奇异值分布较平坦，如果仅用一两个奇异值可能造成信息丢失。通常奇异值数量选取方法有比例法、转折点法、阈值法、累积相对方差法。本文用累积相对方差控制奇异值数量 l，与前三种方法相比

较，累积相对方差能保留大多有用的信息。设定阈值 σ，使得前 l 个奇异值的累积方差和与所有奇异值累积方差和之比大于阈值 σ：$\mathrm{argmin}\left(l:\sum_{k=1}^{l}\delta_k \geqslant \sigma\right)$。阈值 σ 不同时，得到的奇异值数量也不同，数据分析结果也不同。最佳阈值 σ_{opt} 是数据分析结果达到最优时的值。当数据源不同时，最佳阈值变化范围较大，应采用多次实验的方法来选取最优值，基因表达数据一般取 $\sigma=0.7\sim0.9$[10]，选定阈值 σ 和奇异值数量 l，得到近似矩阵 $\boldsymbol{X}^{(l)}$

$$\boldsymbol{X}_{m\times n}^{(l)}=\boldsymbol{U}_{m\times l}^{(l)}\boldsymbol{S}_{l\times l}^{(l)}\boldsymbol{V}_{l\times n}^{(l)\mathrm{T}} \tag{3}$$

1.3　Logistic 回归模型

多数基于 SVD 的基因表达数据分析是时间序列数据[1-4]，得到的特征模式代表包含在原数据矩阵中的基因表达模式，特征模式按对应方差大小排列，方差较大的特征模式往往包含更多的信息。但是在分析样本分类或特征基因选择时，仍用方差评判特征模式对分类作用的大小会产生偏差，方差大的特征模式不一定对分类的贡献就大，因为方差是特征模式在所有样本中的方差，它忽略了样本的类别标识。因此，方差已不适于评估特征模式的分类作用大小。本算法用 Logistic 回归评估每个特征模式对分类贡献大小。

Logistic 回归属于概率型非线性回归，是分析因变量为分类变量而非连续变量时常用的统计方法，在医学研究各个领域被广泛应用如流行病学、临床诊断判别模型等[11]。Logistic 回归模型有二分类和多分类，LRSVD 算法中用到的是二分类回归模型，即分类样本属于正常样本或疾病样本，通过 Logistic 曲线来预测因变量所属样本的概率，曲线输出表示属于不同类别的概率。对特征模式矩阵 $\boldsymbol{V}^{(l)\mathrm{T}}$ 进行 Logistic 回归分析，$\boldsymbol{V}^{(l)\mathrm{T}}$ 矩阵的行 $\boldsymbol{v}_k$ 称为解释变量，表示每一条特征模式在不同样本中的表达水平；矩阵的列 $\boldsymbol{a}_j=[v_{1j},v_{2j},\cdots,v_{lj}]^{\mathrm{T}}$，$j=1,2,\cdots,n$ 是某一样本条件下特征基因的表达水平，样本类标 $\boldsymbol{y}=[y_1,y_2,\cdots,y_n]$ 是二进制向量，表示该样本所属的类别，通常 1 表示疾病，0 表示正常。Logistic 预测模型 $\pi_j=P(y_j=1\,|\,\boldsymbol{V}^{(l)\mathrm{T}})$ 通过式(4)建立

$$\mathrm{log}it(\pi_j)=\log\frac{\pi_j}{1-\pi_j}=\beta_0+\sum_{k=1}^{l}\beta_k v_{kj}=\beta_0+\beta_1 v_{1j}+\cdots+\beta_l v_{lj} \tag{4}$$

式中：β_0 表示偏移量；$\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_l]$ 是特征模式回归系数，回归系数绝对值表明特征模式对样本分类的贡献大小，回归系数绝对值大，该特征模式对样本分类影响大，回归系数接近零，这个特征模式对样本分类没有什么影响。Logistic 回归系数的特性使其比方差更适于评估特征模式的分类能力，用 LRSVD 算法选出的特征模式在不同样本中有很高的差异表达程度。

1.4　基因内积的定义

特征基因选取是在基因表达数据中选取具有高分类准确率的基因子集，所以不能仅对特征模式的分类能力进行评估，需要从特征模式空间转换到原基因表达数据中来选取特征基因。LRSVD 算法通过基因内积实现转换，并按内积大小对基因排序。选取阈值 σ，确定特征模式数量 l，特征模式矩阵 $\boldsymbol{V}^{(l)\mathrm{T}}$ 中的 l 个特征模式形成 l 维的正交坐标系，基因系数矩阵 $\boldsymbol{U}^{(l)}$ 中的每一行表示所对应的基因在特征模式形成的 l 维正交坐标系中的坐标，如图 1 所示，$\boldsymbol{g}_i$ 表示第 i 条基因在 l 维正交坐标系中的坐标值，每个特征模式所对应的坐标轴权重用 Logistic 回归系数表示，权重大的特征模式对分类作用大，如果第 i 条基因在该坐标轴的坐标值大，说明第 i 条基因在该坐标轴的投影也较大。

内积 IP_i(Inner-product)用于评估每一条基因的分类能力，其定义是绝对坐标值向量与绝对回归系数向量的内积，定义如下

$$IP_i=\mathrm{abs}(\boldsymbol{g}_i)\cdot\mathrm{abs}(\boldsymbol{\beta})\quad i=1,2,\cdots,m \tag{5}$$

如果 IP_i 值比较大，说明第 i 条基因在分类能力强的特征模式方向投影大，所以第 i 条基因的分类能力也比较强。

2　LRSVD 算法流程

LRSVD 算法按内积大小对基因排序并删除排序低的基因，用支持向量机(Support vector machine, SVM)评估分类准确率[12]，分类准确率最高的基因子集即是最优特征基因子集。

LRSVD 算法详细步骤如下：

(1) 对预处理后的原始基因表达数据矩阵 $\boldsymbol{X}_{m\times n}$ 作奇异值分解，得到奇异向量 $\boldsymbol{u}_k$，$\boldsymbol{v}_k$ 和奇异值 s_k。

(2) 选定阈值 σ，计算最小特征模式数量 l。

(3) 用 Logistic 回归方法计算 l 个特征模式回归系数 $\boldsymbol{\beta}$；按式(5)计算每条基因的 IP_i 值并按降序排列。

(4) 按比例删除排序较低的部分基因，删除比例一般是 10%～20%。

(5) 用 SVM 评估剩余基因子集的分类能力，分类能力用分类准确率(Accuracy)表示

$$\mathrm{Accuracy}=(TP+TN)/(TP+FP+TN+FN) \tag{6}$$

式中：TP，FP，TN，FN 分别表示正阳性、负阳性、正阴性、负阴性。

(6) 重复步骤(1～5),分类算法准确率最高的基因子集即为本次迭代最佳特征基因子集。

(7) 选择不同的阈值 σ 重复进行上述步骤,直到选出最优的基因子集和最佳阈值。

3 实验结果及分析

将 LRSVD 算法应用于实际基因表达数据 Colon 数据[13]和前列腺癌数据[14],SVM 采用径向基核函数(Radial basis function,RBF),实验结果采用十倍交叉验证。

3.1 Colon 数据实验结果

结肠癌数据(Colon data)是二分类肿瘤样本数据,共有 2 000 条基因,62 个样本,其中 22 个正常样本,40 个疾病样本,首先将基因数据标准化,每条基因平均表达强度为 0,标准方差为 1。选择不同的 σ 值:0.45,0.65,0.85,0.90,每一个 σ 值产生一系列基因子集,用 SVM 评估得到的基因子集,分类准确率最高且基因数量最少的一组基因为最佳特征基因子集。从图 2 可知,当 $\sigma=0.85$ 时,得到最高分类准确率 0.935 5,最佳分类基因子集中基因数量是 21。图 3 是这 21 条基因在两类样本的表达谱,前 22 个是正常样本,后 40 个是疾病样本,白色代表最大表达值,黑色代表最小表达值,这 21 条基因在两类样本中差异表达,表 1 列出了这 21 条基因在数据库中的基因登录号和基因描述。

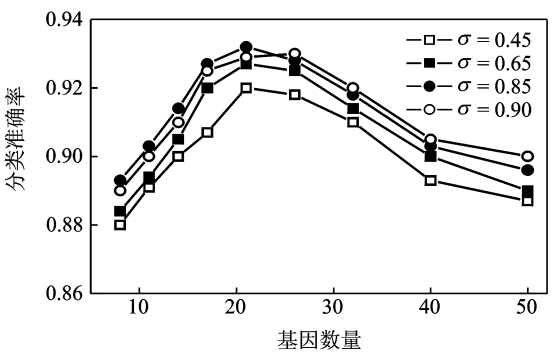

图 2 参数 σ 取不同值的分类准确率

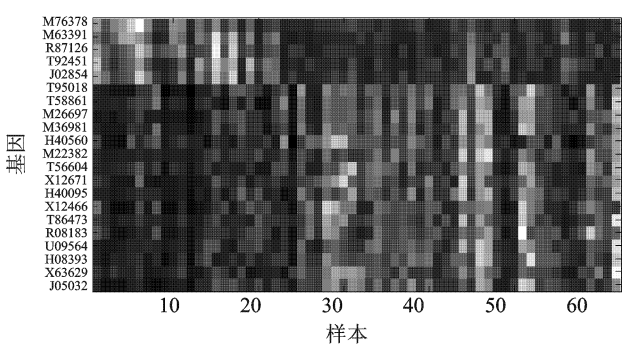

图 3 结肠数据最佳基因子集在两类样本中的表达谱(22 正常样本/40 疾病样本)

3.2 前列腺癌数据实验结果

前列腺癌数据是由寡核苷酸芯片得到的数据,包括 9 个正常样本,25 个疾病样本,大约 12 600 条基因和 ESTs 序列。首先对数据进行阈值替换、基因过滤和对数转换等预处理,将 LRSVD 算法应用于预处理后的 2 000 条基因数据,阈值参数 σ 分别选取不同的值:0.45,0.65,0.85,0.90。当 $\sigma=0.85$ 时,得到有 5 条基因的最佳基因子集,最高分类准确率 97.06%,即在 34 个样本中有一个样本错分。表 2 列出了这 5 条基因的描述,图 4 是这 5 条基因在两类样本中的表达谱,白色代表最大表达值,黑色代表最小表达值,图中前 9 个是正常样本,后 25 个是疾病样本,这 5 条基因在两类样本中差异表达。

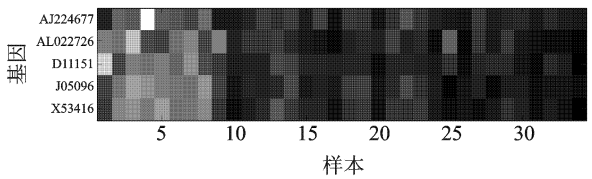

图 4 前列腺癌数据最佳基因子集在两类样本中的表达谱

3.3 LRSVD 算法实验结果分析

将结肠癌基因表达数据实验结果与其他研究者的结果进行比较,LRSVD 算法取得 93.55% 的准确率,是结肠癌数据能达到的较好分类结果。LRSVD 算法挖掘的特征基因与大多数的研究结果相吻合,Wessels 等研究者用两种 U-检验法和互信息作为过滤标准对结肠数据中的基因进行排序,选取了综合排序较高的 7 条基因,这 7 条基因中有 6 条与表 1 中的结果相吻合[15]。Shaik 等人用 unified framework 方法从结肠数据集中选出了 66 个差异表达的基因,这 66 条基因包含了表 1 中的大多数基因[16]。Bicciato 等人用自关联神经网络方法挖掘出的基因也与本文中选择的约半数结肠数据特征基因相一致[17]。Xiong 等人用线性判别分析、Logistic 回归、支持向量机方法作为分类算法选择的基因包含有 M63391,R87126,T92451,J02854,H08393,J05032,M76378[18]。Diao 等人用贝叶斯网络聚类方法挖掘疾病基因之间的相关性,聚类结果中包含有基因 R87126,J02854,M63391,M76378,T92451 等[19]。Li 等人用集成决策方法挖掘的致病基因中包含有 M63391,R87126,H08393,J02854 等基因[20],其他结肠数据致病基因挖掘结果也与本文研究结果部

分一致[21]。

除分类准确率外,特征基因的生物功能是更应该关注的问题,这些基因是否与疾病相关,以及在疾病发生过程中的生物机制。本文从多个生物数据库对表 1 的基因进行生物功能查询,发现这些基因具有较强的生物意义,并与疾病相关。

表 1　结肠癌数据中最佳特征基因子集

基因登录号	基因描述
T95018	40S ribosomal protein S18 S18 (Homo sapiens)
T58861	60S RIBOSOMAL PROTEIN L30E (Kluyveromyces lactis)
M26697	Human nucleolar protein (B23) mRNA, NPM1
M36981	Human putative NDP kinase (nm23-H2S) mRNA, NME2
M76378	Human cysteine-rich protein (CRP) gene
M63391	Human desmin gene, DES
H40560	Thioredoxin(Human)
R87126	Myosin heavy chain, nonmuscle (Gallus gallus)
M22382	Mitochondrial matrix protein P1 precursor(Human),HSPD1
T56604	Tubulin beta chain (Haliotis discus)
X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
H40095	Macrophage migration inhibitory factor (Human)
T92451	Tropomyosin, fibroblast and epithelial muscle-type (Human), TPM2
X12466	Human mRNA for snRNP E protein
T86473	Nucleoside diphosphate kinase A (Human)
R08183	10 kd heat shock protein, mitochondrial
U09564	Human serine kinase mRNA
J02854	Myosin regulatory light chain 2, smooth muscle isoform (Human); MYL2
H08393	Collagen alpha 2 (XI) chain (Human)
X63629	H. sapiens mRNA for p cadherin; CDH3
J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA

表 2　前列腺癌数据中最佳特征基因子集

探针 ID	基因登录号	基因描述
35837_at	AJ224677	Homo sapiens mRNA for scrapie responsive protein 1, SCRG1
41536_at	AL022726	Inhibitor of DNA binding 4, dominant negative helix-loop-helix protein, ID4
1507_s_at	D11151	Endothelin-A receptor
34377_at	J05096	Na,K-ATPase subunit alpha 2 gene, ATP1A2
32750_r_at	X53416	Human mRNA for actin-binding protein (filamin), FLNA

4　结束语

从本文实验结果可得出结论,LRSVD 算法能有效地挖掘出与疾病相关的基因子集,得到较高的分类准确率,挖掘出的特征基因具有生物意义,有助于解释基因功能及疾病的发病机理。LRSVD 算法通过阈值 σ 控制累积方差和,进而控制特征基因数量,但是关于最佳阈值 σ 并没有太多理论研究,对于给定数据,通过分析数据来源、数据类型、噪声含量,大多采取多次尝试法获得最佳阈值分布范围。本文后续工作中将运用优化方法对阈值 σ 的选取进行理论研究,进一步完善 LRSVD 算法在特征基因选择方面的应用。

参考文献:

[1] Alter O, Brown P O, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling[J]. Proc Natl Acad Sci USA, 2000,97(18):10101-10106.

[2] Berrar D P, Dubitzky W, Granzow M. A practical approach to microarray data analysis[M]. Kluwer: Norwell, MA, 2003:91-109.

[3] Cho R J, Campbell M J, Winzeler E A, et al. A genome-wide transcriptional analysis of the mitotic cell cycle[J]. Mol Cell, 1998,2(1):65-73.

[4] Holter N S, Maritan A, Cieplak M, et al. Dynamic modeling of gene expression data[J]. Proc Natl Acad

Sci, 2001,98(4):1693-1698.

[5] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays [J]. Bioinformatics, 2001,17(6):520-525.

[6] Wall M E, Dyck P A, Brettin T S. SVDMAN——singular value decomposition analysis of microarray data[J]. Bioinformatics, 2001,17(6):566-568.

[7] Hastie T, Tibshirani R, Eisen M B, et al. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns[J]. Genome Biol, 2000,1(2):Research0003:1-21.

[8] Sill M, Kaiser S, Benner A, et al. Robust biclustering by sparse singular value decomposition incorporating stability selection[J]. Bioinformatics, 2011, 27 (15):2089-2097.

[9] Bécavin C, Tchitchek N, Mintsa-Eya C, et al. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition [J]. Bioinformatics, 2011, 27 (10):1413-1421.

[10] Liang F M. Use of SVD-based probit transformation in clustering gene expression profiles[J]. Computational Statistics & Data Analysis, 2007, 51 (12): 6355-6366.

[11] Liao J G, Khew-Voon C. Logistic regression for disease classification using microarray data: model selection in a large p and small n case[J]. Bioinformatics, 2007,23(15):1945-1951.

[12] Lee Y, Lee C K. Classification of multiple cancer types by multicategory support vector machines using gene expression data[J]. Bioinformatics, 2003, 19 (9):1132-1139.

[13] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences, 1999,96(12):6745-6750.

[14] Welsh J B, Sapinoso L M, Su A I, et al. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer[J]. Cancer Res, 2001,61(16):5974-5978.

[15] Wessels L F A, Reinders M J T, Nederlof P M, et al. Representation and classification for high-throughput data sets[C]// SPIE-BIOS2002, Biomedical Nanotechnology Architectures and Applications. USA:[s. n.], 2002, 4626:226-237.

[16] Shaik J, Yeasin M. A unified framework for finding differentially expressed genes from microarray experiments[J]. BMC Bioinformatics, 2007,8:347.

[17] Bicciato S, Pandin M, Didonè G, et al. Pattern identification and classification in gene expression data using an autoassociative neural network model[J]. Biotechnol Bioeng, 2003,81(5):594-606.

[18] Xiong M M, Fang X Z, Zhao J Y. Biomarker identification by feature wrappers[J]. Genome Research, 2001,11(11):1878-1887.

[19] Diao Q, Hu W, Zhong H, et al. Disease gene explorer: display disease gene dependency by combining Bayesian networks with clustering[C]// Proceedings of IEEE Computational Systems Bioinformatics Conference. [S. l.]:IEEE, 2004:574-575.

[20] Li X, Rao S, Zhang T, et al. Gene mining:a novel and powerful ensemble decision approach to hunting for genes using microarray expression profiling[J]. Nucleic Acids Research, 2004,32(9):2685-2694.

[21] Zhang X W, Yap Y L, Wei D, et al. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis[J]. European Journal of Human Genetics, 2005, 13 (12): 1303-1311.